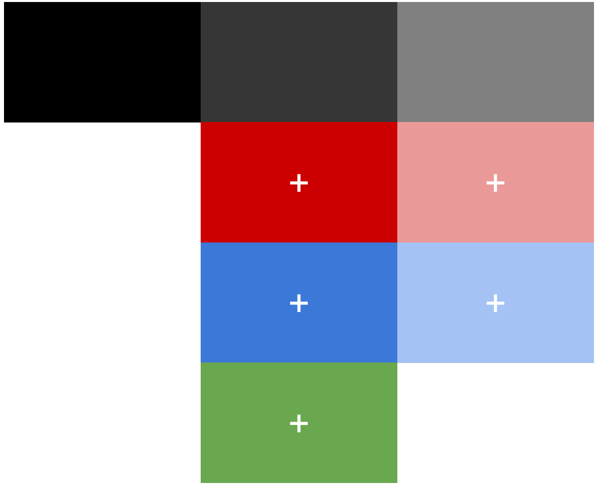
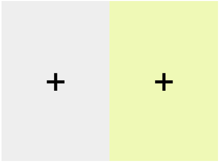
Color check

if this is unreadable, we’re in trouble

if this is unreadable, whatever





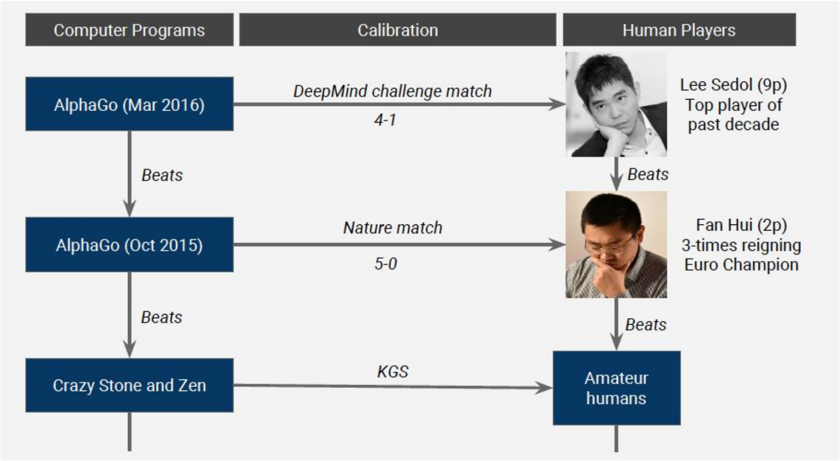
Adversarial Example Defenses:

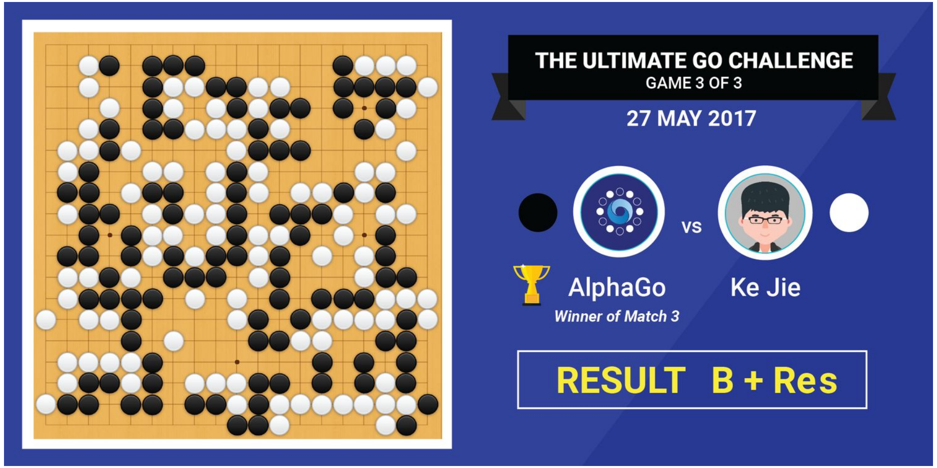
Ensembles of Weak Defenses are not Strong

Warren He
James Wei
Xinyun Chen
Nicholas Carlini
Dawn Song

UC Berkeley

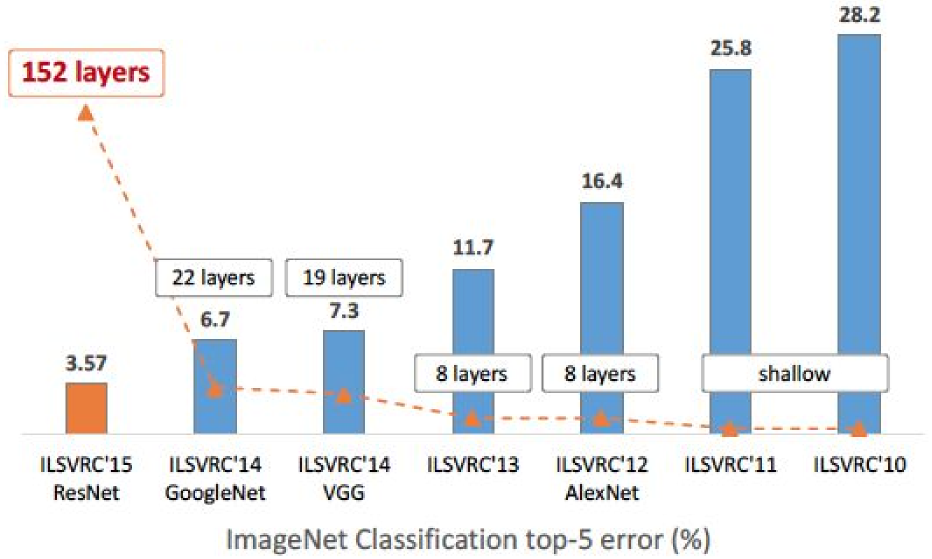
AlphaGo: Winning over World Champion





Source: David Silver

Achieving Human-Level Performance on ImageNet Classification



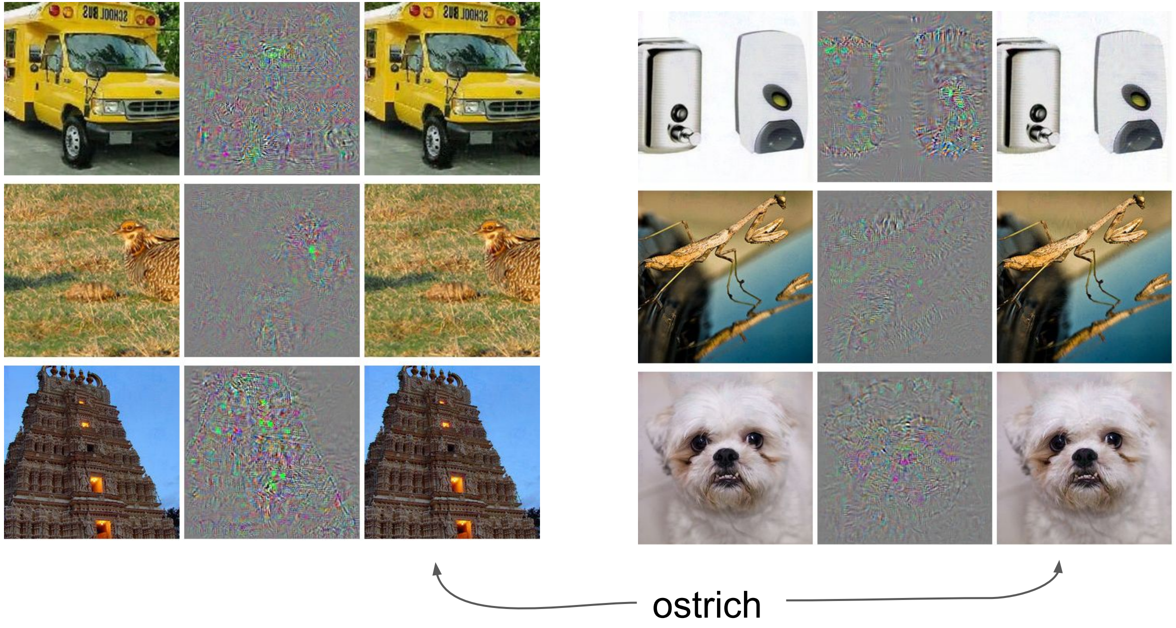
Source: Kaiming He

Deep Learning Powering Everyday Products

pcmag.com

theverge.com

Deep Learning Systems Are Easily Fooled



$$\frac{\partial \text{output}}{\partial \text{pixels}}$$

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. Intriguing properties of neural networks. ICLR 2014.

Outline

Background: neural networks and adversarial examples

Defenses against adversarial examples

Ensemble defenses case studies

- Feature squeezing
- Specialists+1
- Unrelated detectors

Conclusion

Outline

Background: neural networks and adversarial examples

Defenses against adversarial examples

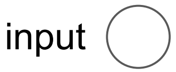
Ensemble defenses case studies

- Feature squeezing
- Specialists+1
- Unrelated detectors

Conclusion

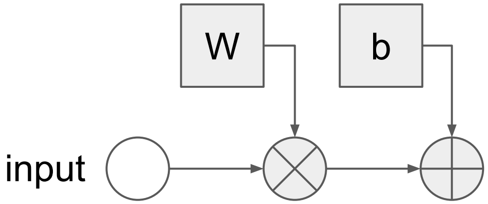
Background: Neural networks

Input: a vector of numbers, e.g., image pixels



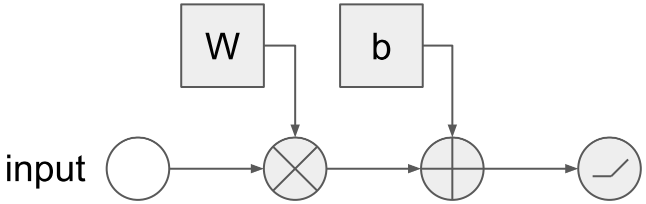
Background: Neural networks

Linear combination (matrix multiply) and add bias



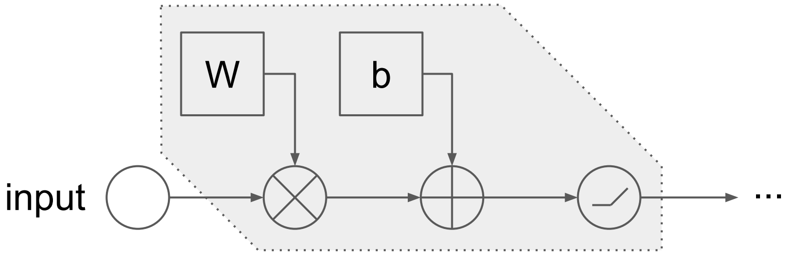
Background: Neural networks

Nonlinearity, e.g., ReLU(x) = max(0, x)



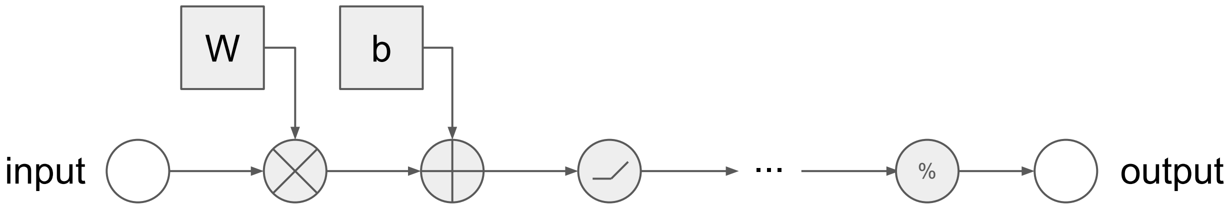
Background: Neural networks

Many layers of these (deep)



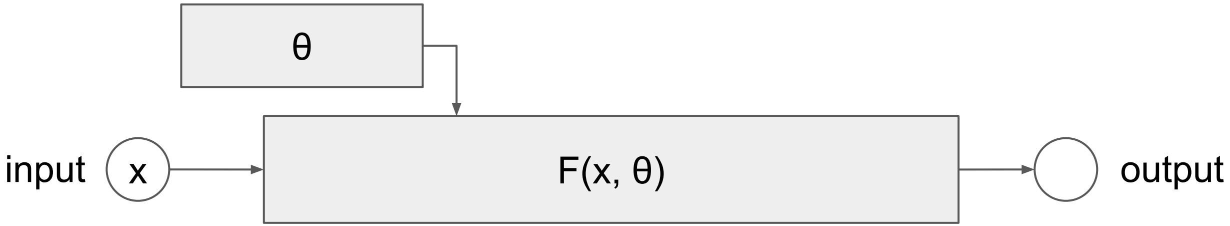
Background: Neural networks

In image classification, *softmax* function converts output to probabilities



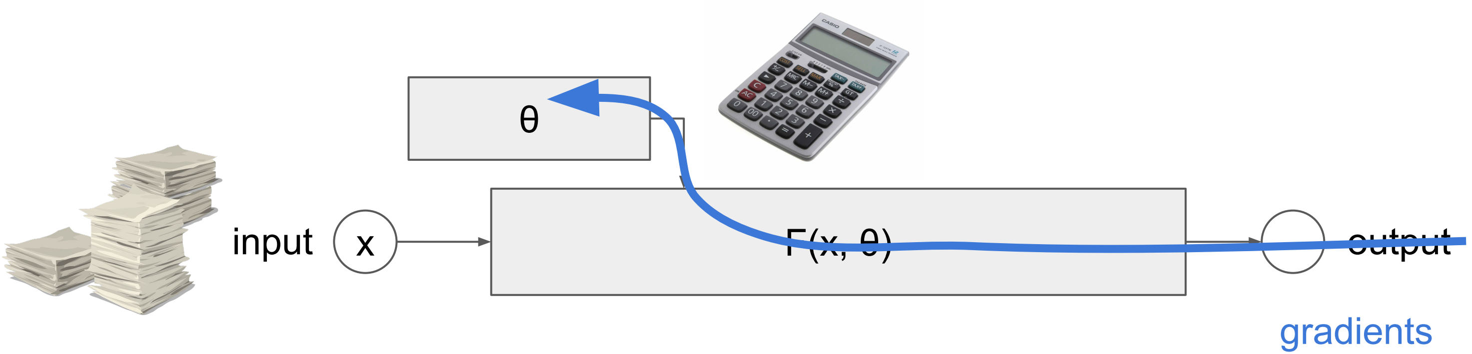
Background: Neural networks

Overall, a great big function that takes an input **x** and parameters **θ**.



Background: Neural networks

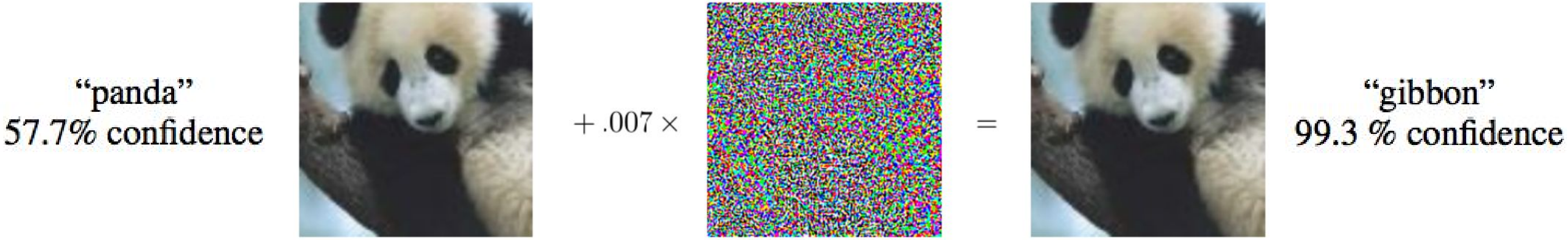
Overall, a great big function that takes an input **x** and parameters **θ**.



Some training data in **x**, know what output should be,
use **gradient descent** to figure out best **θ**.

Background: Adversarial examples

Small change in input, wrong output.



Smallness referred to as **distortion**.

Measured in L_2 distance:
Euclidean distance if image were a vector of pixel values

Background: Adversarial examples

~~State of the art~~: Vulnerable

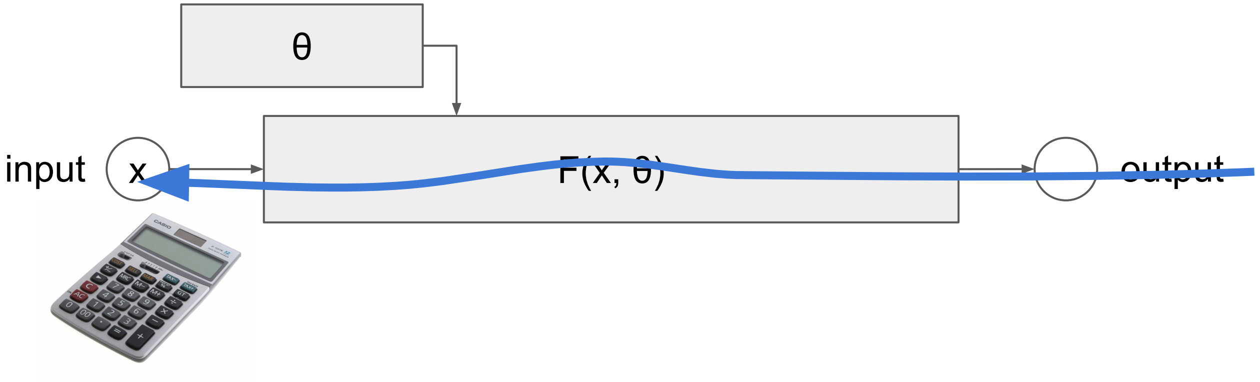
- Image classification
- Caption generation
- Speech recognition
- Natural language processing
- Policies, reinforcement learning
- Self-driving cars

Stop Sign

Yield Sign

Background: Generating adversarial examples

How? Gradients again.



Differentiate with respect to inputs, rather than parameters

Get: how to change each pixel to make output a little more wrong

Background: Generating adversarial examples

We have gradient → We optimize

Given original input x and correct output y:

$$\min_{x'} ||x' - x||_2^2 \text{ s.t. } F(x', \theta) \neq y$$

close to original

where output is wrong

some other input

Background: Other threat models

These were **white-box** attacks, where attacker knows the model parameters.

Black-box scenarios have less information available.
There are techniques to use white-box attacks in black-box scenarios.

We focus on white-box attacks in this work.

Outline

Background: neural networks and adversarial examples

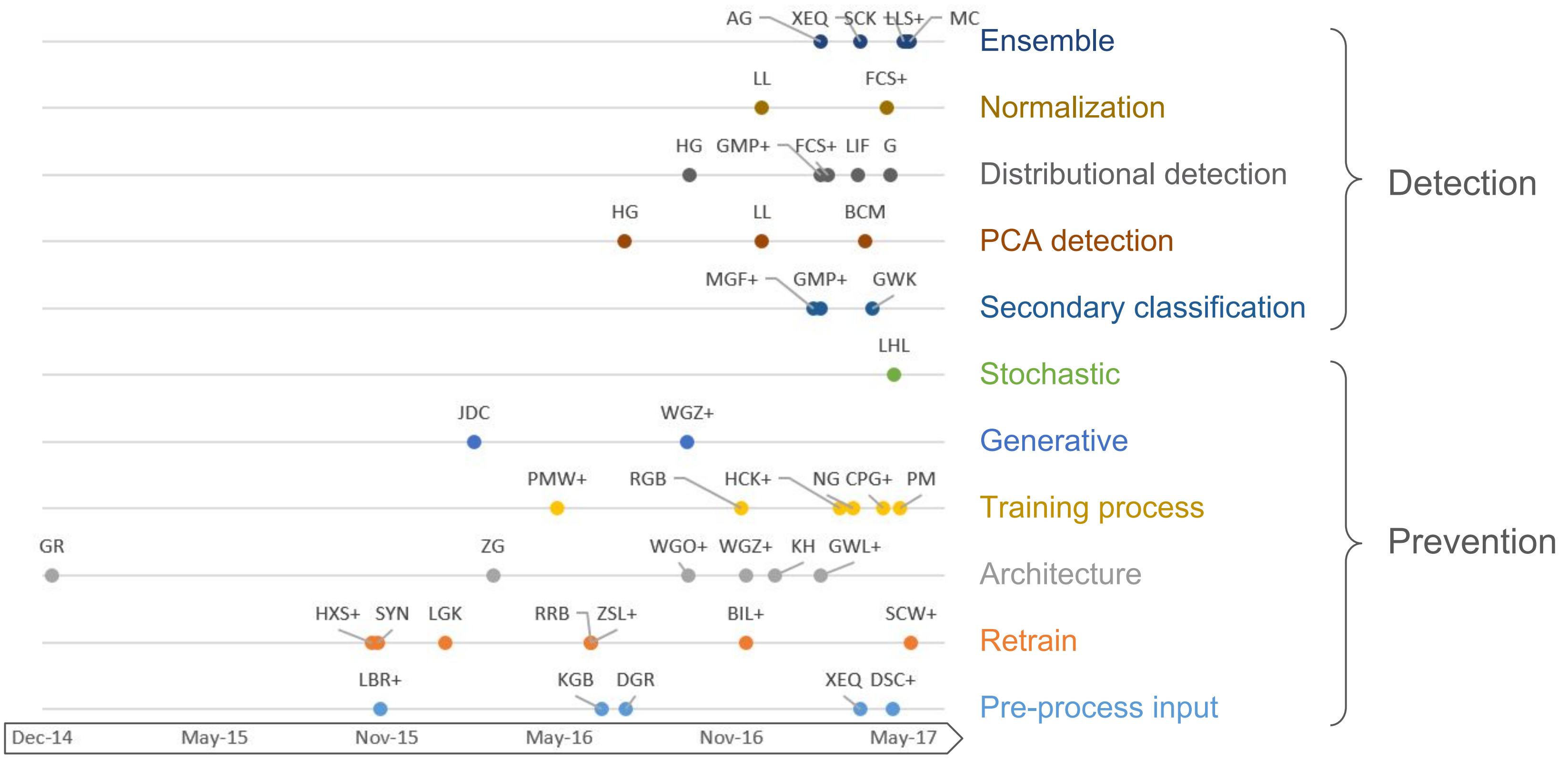
Defenses against adversarial examples

Ensemble defenses case studies

- Feature squeezing
- Specialists+1
- Unrelated detectors

Conclusion

Background: Defenses



Background: Defenses

We evaluate defenses:

- Can we still algorithmically find adversarial examples?
- Do we need higher distortion?

Outline

Background: neural networks and adversarial examples

Defenses against adversarial examples

Ensemble defenses case studies

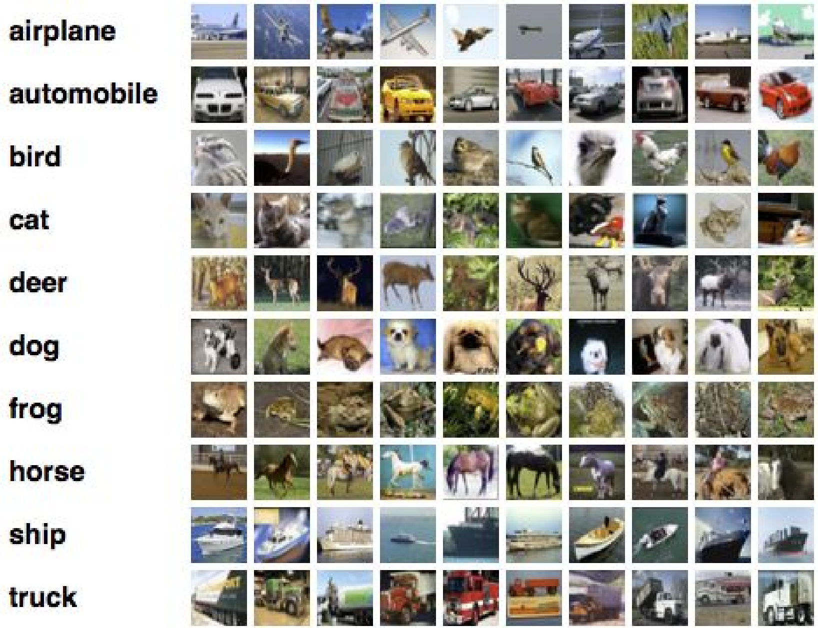
- Feature squeezing
- Specialists+1
- Unrelated detectors

Conclusion

Data sets

MNIST

CIFAR-10



Are ensemble defenses stronger?

Stronger!	Not much stronger

Outline

Background: neural networks and adversarial examples

Defenses against adversarial examples

Ensemble defenses case studies

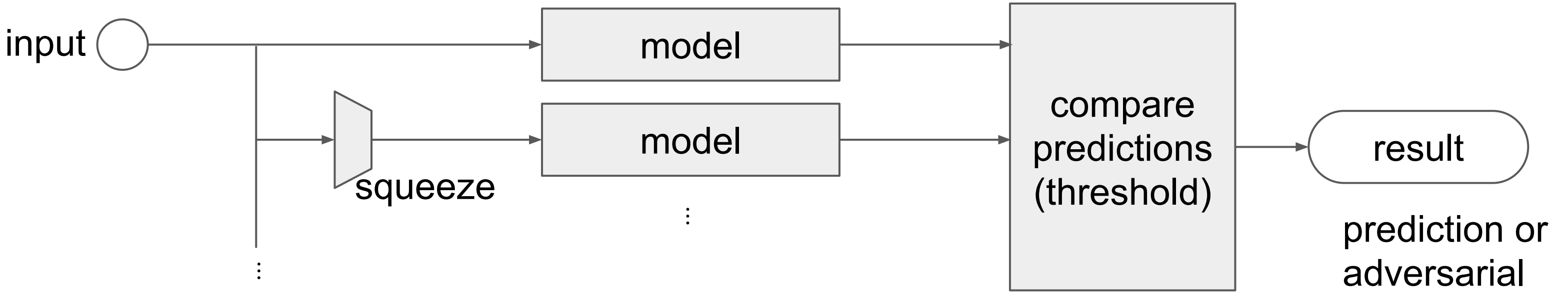
- **Feature squeezing**
  Address two kinds of perturbations
- Specialists+1
- Unrelated detectors

Conclusion

Ensemble defense: Feature squeezing

Run prediction on multiple versions of an input image

Use “squeezing” algorithms to produce different versions of input



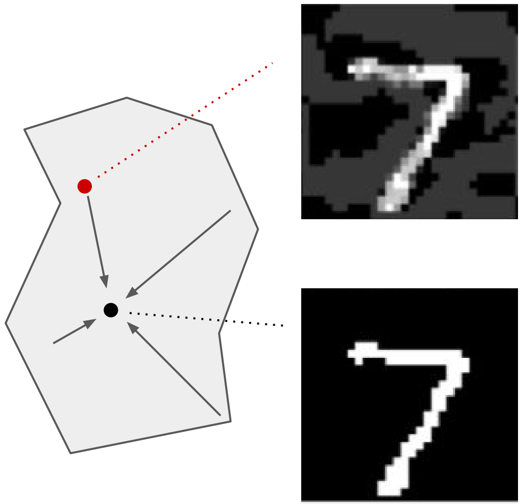
If predictions differ too much, input is adversarial

Xu, W., Evans, D., & Qi, Y. (2017). Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. arXiv preprint arXiv:1704.01155.

Ensemble defense: Feature squeezing

"Squeezing" an image removes some of its information

Maps many images to the same image:
Ideally maps adversarial examples
to something easier to classify



Feature squeezing algorithms and attacks

Two specific squeezing algorithms

- Color depth reduction
- Spatial smoothing

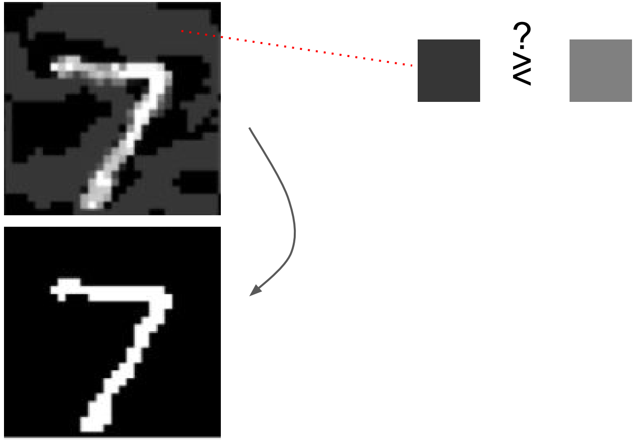
Effectiveness when used in isolation

Squeezing algorithms

Color depth reduction

Convert image colors to low bit-depth

Eliminates small changes
on many pixels



Squeezing algorithms

Color depth reduction

Works well against **fast gradient sign method** (FGSM)

$$x' = x + \epsilon\,\mathrm{sign}(\nabla_x \mathrm{wrongness}(F(x, \theta)))$$

Instead of optimizing, do one quick step

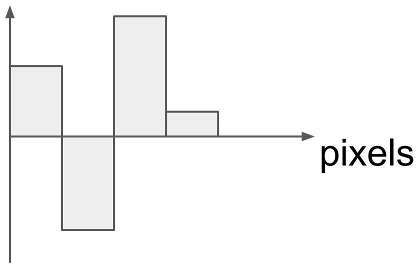
Squeezing algorithms

Color depth reduction

Works well against **fast gradient sign method** (FGSM)

$$x' = x + \epsilon \, \mathrm{sign}(\underline{\nabla_x \mathrm{wrongness}(F(x, \theta))})$$

Gradient in direction of wrong prediction, as usual



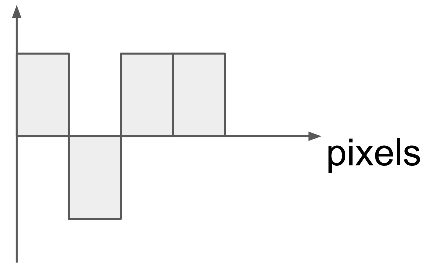
Squeezing algorithms

Color depth reduction

Works well against **fast gradient sign method** (FGSM)

$$x' = x + \epsilon \, \underline{\mathrm{sign}}(\nabla_x \mathrm{wrongness}(F(x, \theta)))$$

Sign of that gradient: only increase or decrease



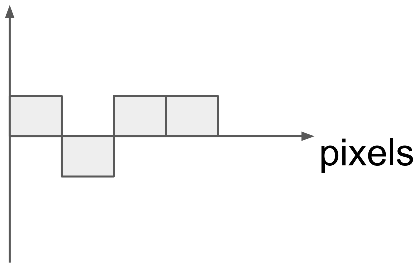
Squeezing algorithms

Color depth reduction

Works well against **fast gradient sign method** (FGSM)

$$x' = x + \epsilon\,\text{sign}(\nabla_x \text{wrongness}(F(x, \theta)))$$

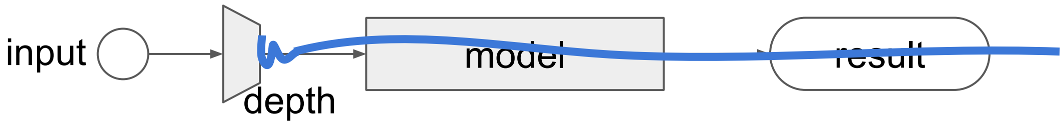
Increase or decrease each pixel by ϵ



Squeezing algorithms

Color depth reduction

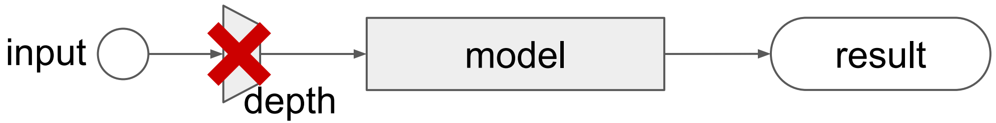
Not fully differentiable



Squeezing algorithms

Color depth reduction

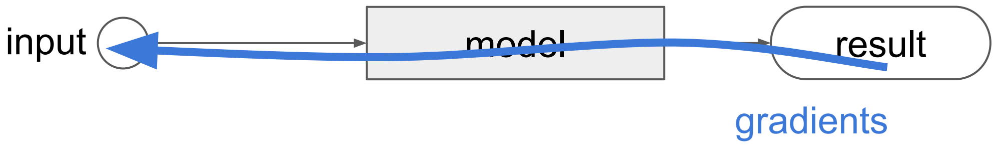
Can be attacked using a substitute that excludes the non-differentiable part



Squeezing algorithms

Color depth reduction

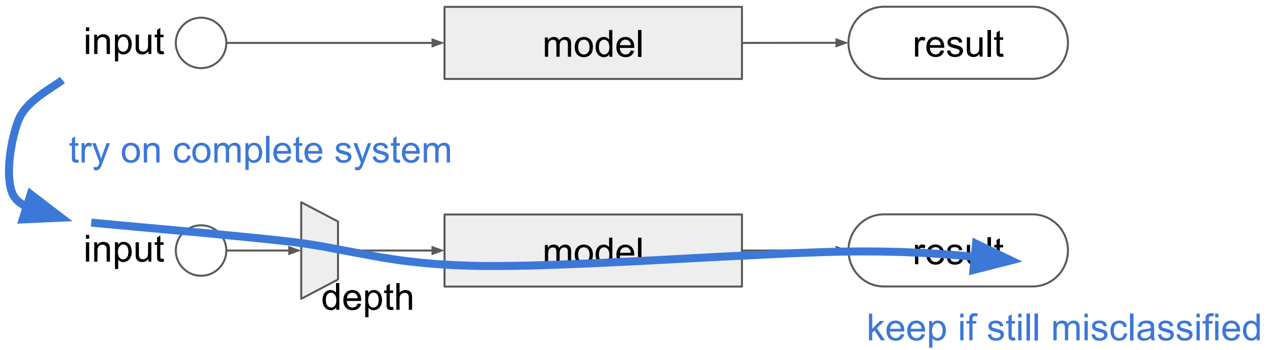
Can be attacked using a substitute that excludes the non-differentiable part



Squeezing algorithms

Color depth reduction

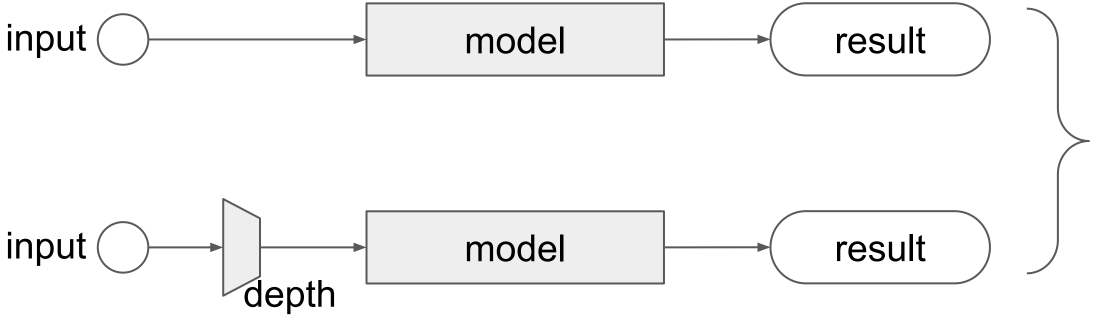
Can be attacked using a substitute that excludes the non-differentiable part



Squeezing algorithms

Color depth reduction

Can be attacked using a substitute that excludes the non-differentiable part



Attacker uses two versions.

One differentiable for generating candidates.

One for testing candidates.

Squeezing algorithms

Color depth reduction: untargeted optimization attack

Can be attacked using a substitute that excludes the non-differentiable part

MNIST, reduction to 1 bit

original

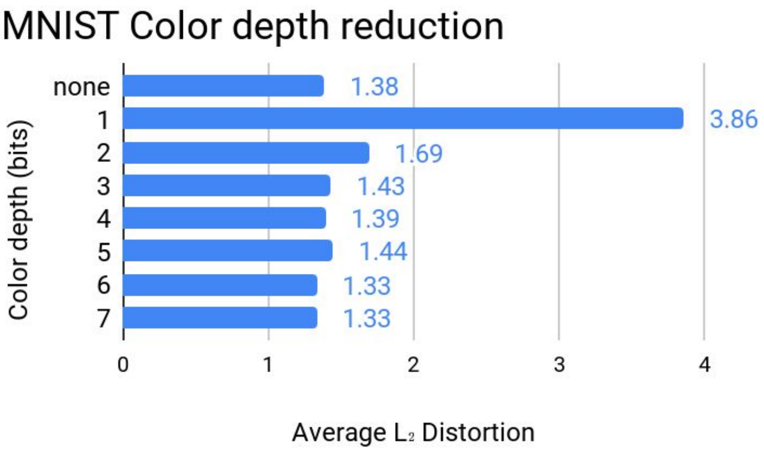
adversarial

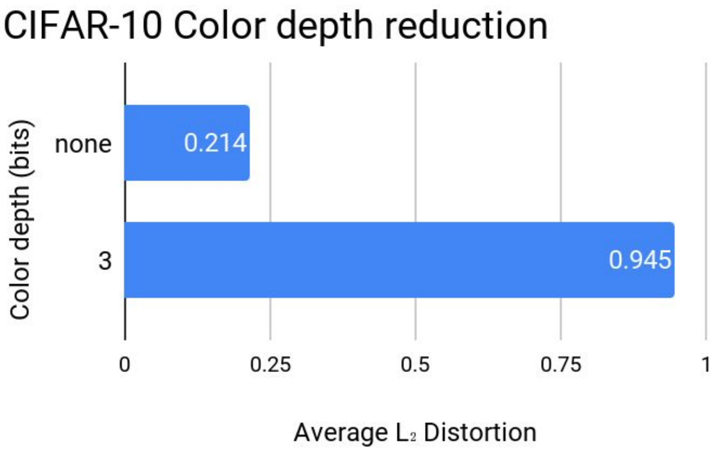
CIFAR-10, reduction to 3 bits

Squeezing algorithms

Color depth reduction

99%-100% success rate, but increases average L_2 distortion



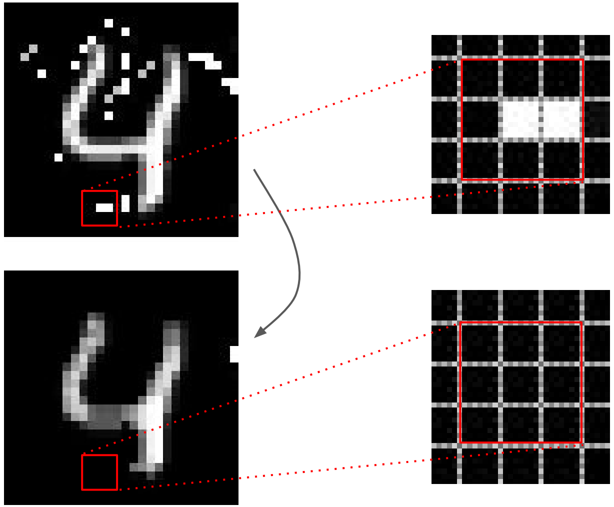


Squeezing algorithms

Spatial smoothing

Median filter:
replace each pixel with median around its neighborhood

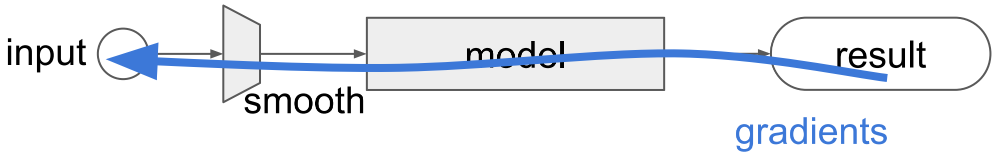
Eliminates strong changes on a few pixels



Squeezing algorithms

Spatial smoothing

Can be attacked directly using existing techniques



Squeezing algorithms

Spatial smoothing: untargeted optimization attack

Can be attacked directly using existing techniques

MNIST, 3×3 smoothing

original

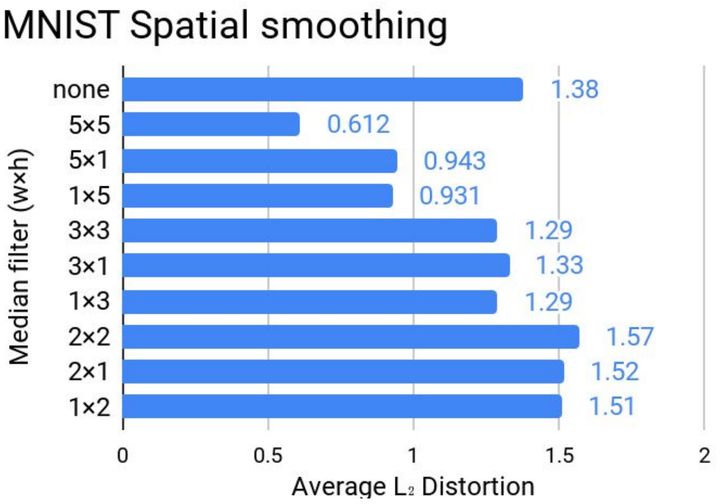
adversarial

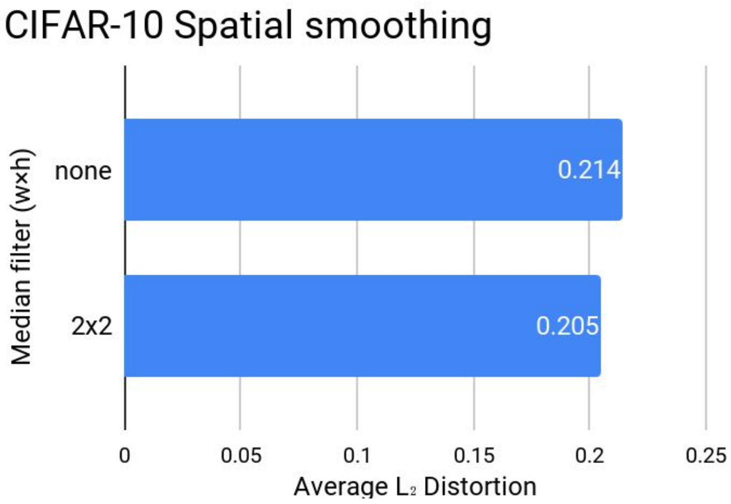
CIFAR-10, 2×2 smoothing

Squeezing algorithms

Spatial smoothing

100% success rate, about the same average L_2 distortion

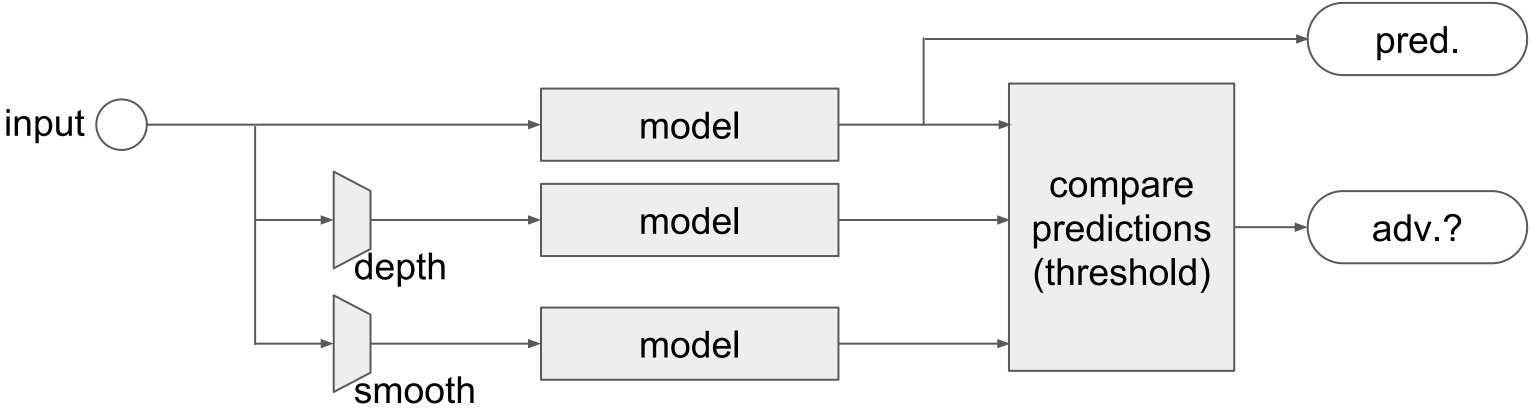




Feature squeezing

Full defense combines these squeezing algorithms in an ensemble.

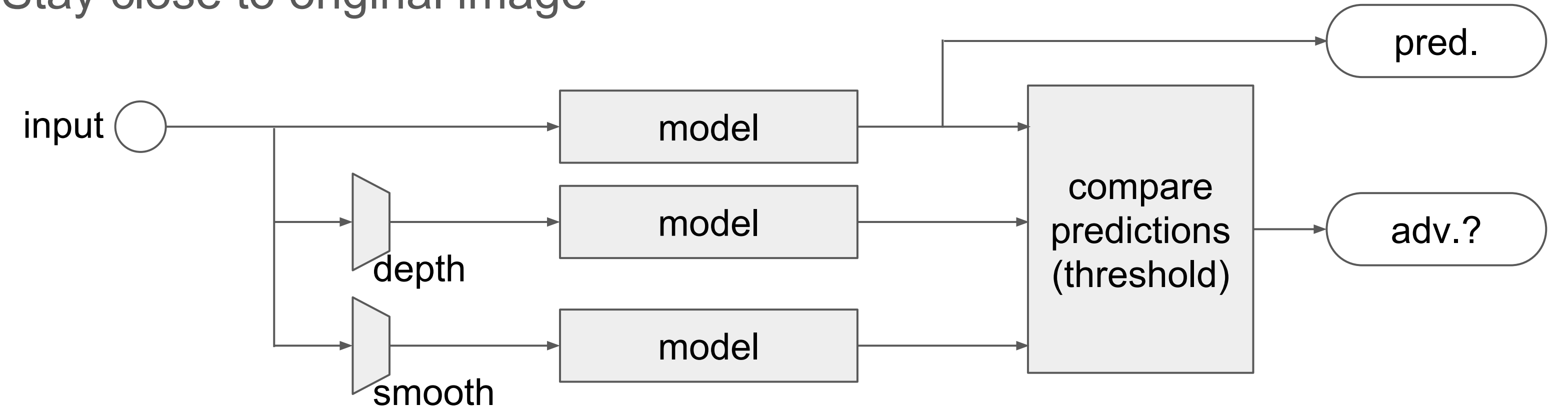
If predictions differ by too much (L_1 distance), input is adversarial.



Feature squeezing: Attack

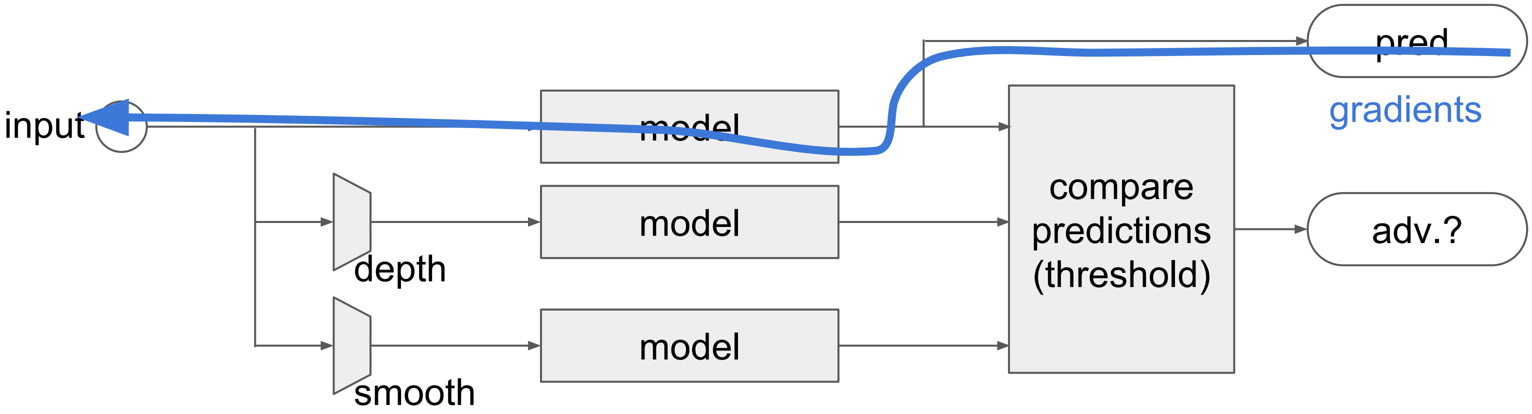
Loss function

- Make prediction wrong
- Make all predictions have low L_1 distance
- Stay close to original image



Feature squeezing: Attack

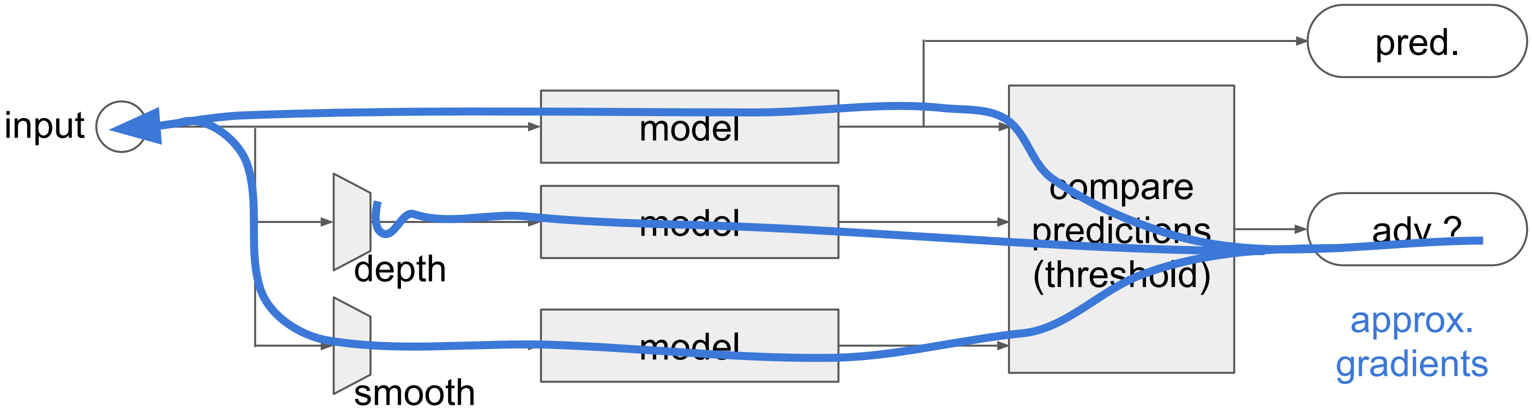
Wrong prediction is fully differentiable



Feature squeezing: Attack

L_1 distance only gets gradients from two branches.

Attacker tests candidates on complete system.



Feature squeezing: Attack

Ensemble defense

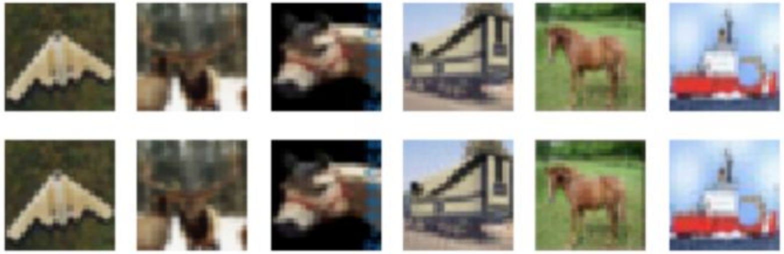
Can be attacked using gradients from differentiable branches and random initialization

MNIST, 1-bit, 3×3

original

adversarial

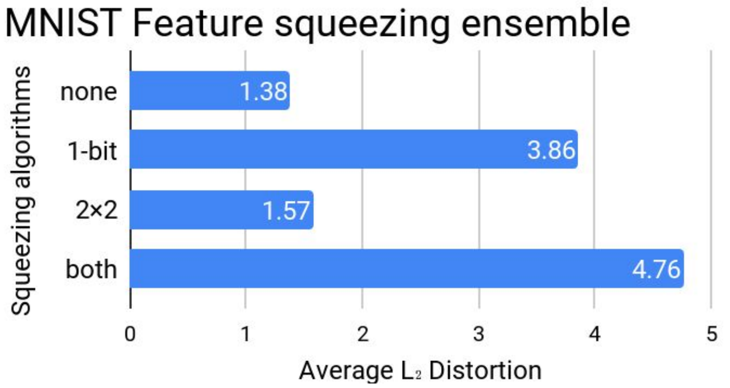
CIFAR-10, 3-bit, 2×2

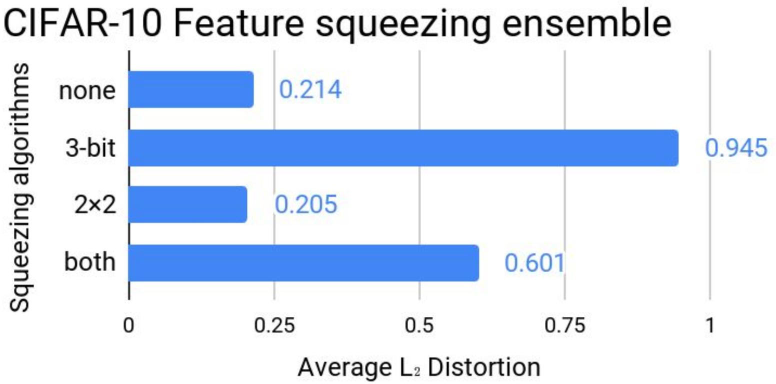


Feature squeezing: Attack

Ensemble defense

100% success rate, average adversarial-ness less than original images, average L_2 distortion not much higher than individual squeezing algorithms

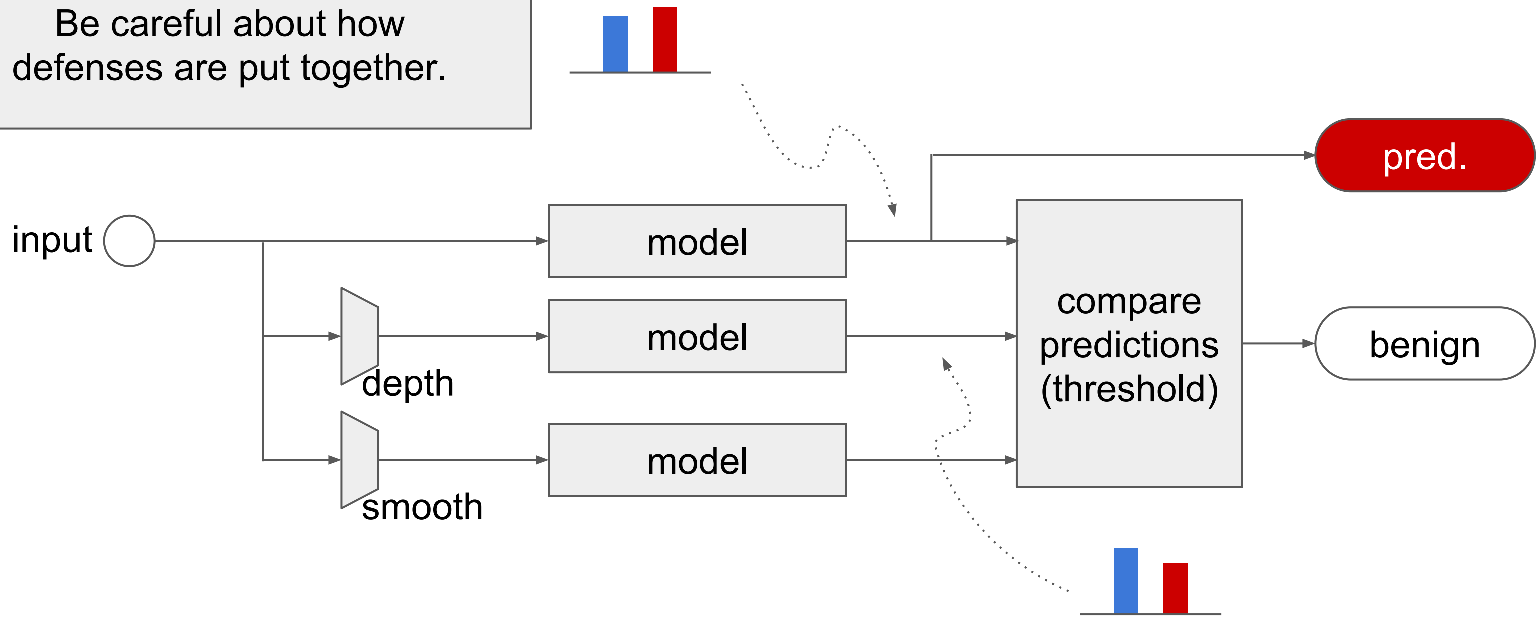




Feature squeezing

Don’t have to completely fool the strongest component defense

Be careful about how defenses are put together.



Are ensemble defenses stronger?

Stronger! 	Not much stronger 

Are ensemble defenses stronger?

Stronger! 	Not much stronger 
	Feature squeezing (MNIST) +23%

Are ensemble defenses stronger?

Stronger!	Not much stronger	Weaker?
	Feature squeezing (MNIST) +23%	
		Feature squeezing (CIFAR-10) −36%

Outline

Background: neural networks and adversarial examples

Defenses against adversarial examples

Ensemble defenses case studies

- Feature squeezing
- **Specialists+1**
  Multiple models, to cover common errors
- Unrelated detectors

Conclusion

Ensemble defense: Specialists+1

Combine *specialist* classifiers that classify among sets of confusing classes.

Example: **automobiles** are more often confused with **trucks** than with **dogs**.

“Automobile” includes sedans, SUVs, things of that sort.
“Truck” includes only big trucks. Neither includes pickup trucks.

	Airplane	Auto	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Auto	2.70		7.20	1.94	5.46	0.06	7.86	0.90	8.50	65.38

Abbasi, M., & Gagné, C. (2017). Robustness to Adversarial Examples through an Ensemble of Specialists. ICLR 2017 Workshop Track.

Ensemble defense: Specialists+1

Two sets corresponding to each class:

- The most common confused classes (top 80%)
- The rest of the classes

For auto: truck, ship, frog and airplane, auto, bird, cat, deer, dog, horse

	Airplane	Auto	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Auto	2.70		7.20	1.94	5.46	0.06	7.86	0.90	8.50	65.38

Additionally, a "generalist" set with all classes

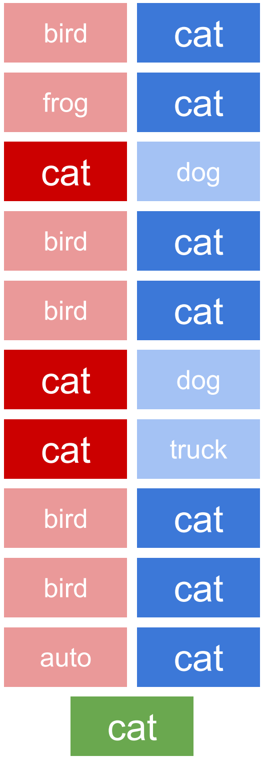
Ensemble defense: Specialists+1

For each set, train a classifier to classify between those classes

If all classifiers that can predict a class do predict that class, then only those classifiers vote; otherwise, all classifiers vote

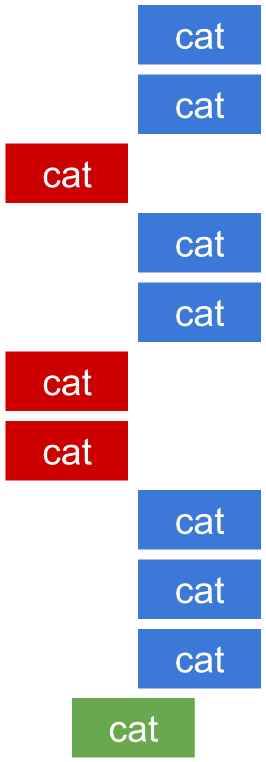
Class with most votes is the prediction

If **average confidence among voting classifiers** is low, then input is adversarial



Specialists+1: attack

Targeted attack: figure out which classifiers would be needed to win with a unanimous vote

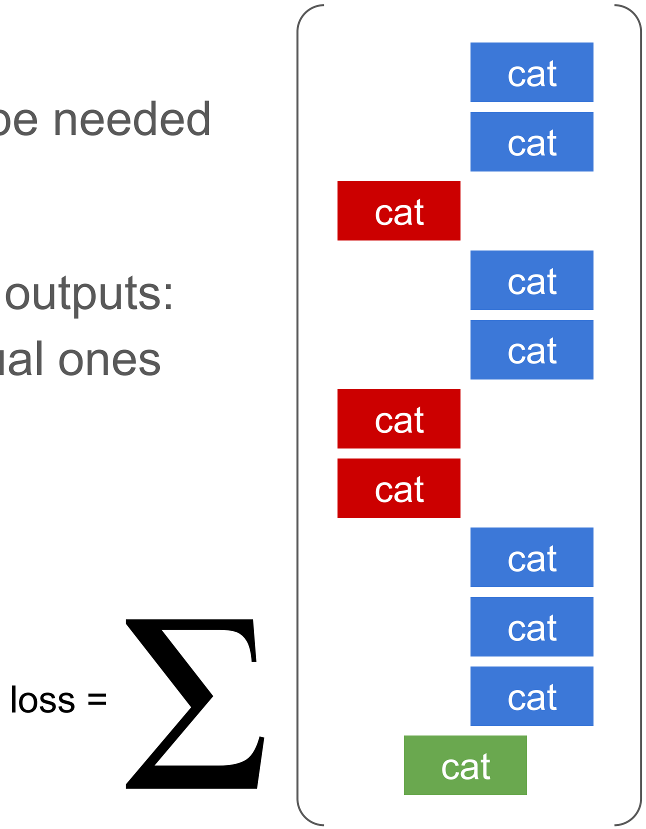


Specialists+1: attack

Targeted attack: figure out which classifiers would be needed to win with a unanimous vote

Optimize loss function made from those classifiers’ outputs: add up loss functions that we would use for individual ones

Favor high confidence, not just misclassification



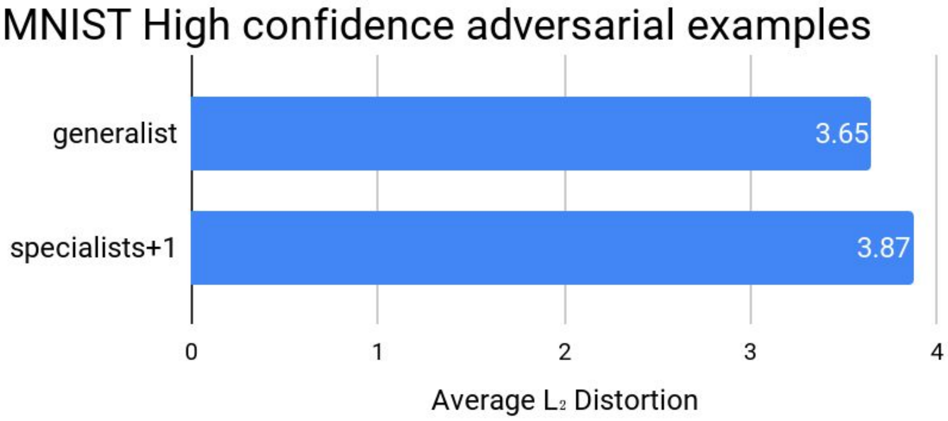
Specialists+1: attack

Targeted optimization attack

Specialists+1: attack

Targeted optimization attack

Randomly chosen targets, 99% success rate,
average confidence higher than average of benign images



Are ensemble defenses stronger?

Stronger! 	Not much stronger 	Weaker? 
	Feature squeezing (MNIST) +23%	
		Feature squeezing (CIFAR-10) −36%
	Specialists+1 (MNIST) +6%	

Outline

Background: neural networks and adversarial examples

Defenses against adversarial examples

Ensemble defenses case studies

- Feature squeezing
- Specialists+1
- **Unrelated detectors**

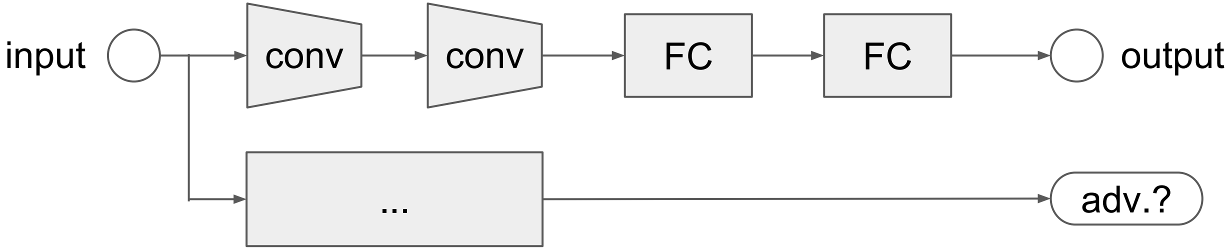
  Does it matter if defenses are designed to work well together?

Conclusion

Three unrelated detectors

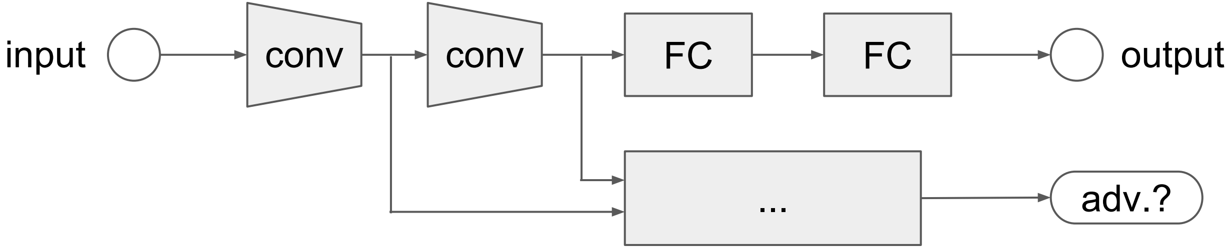
1. A separate network that distinguishes benign and adversarial images.

   GONG, Z., WANG, W., AND KU, W.-S. Adversarial and clean data are not twins. arXiv preprint arXiv:1704.04960 (2017).



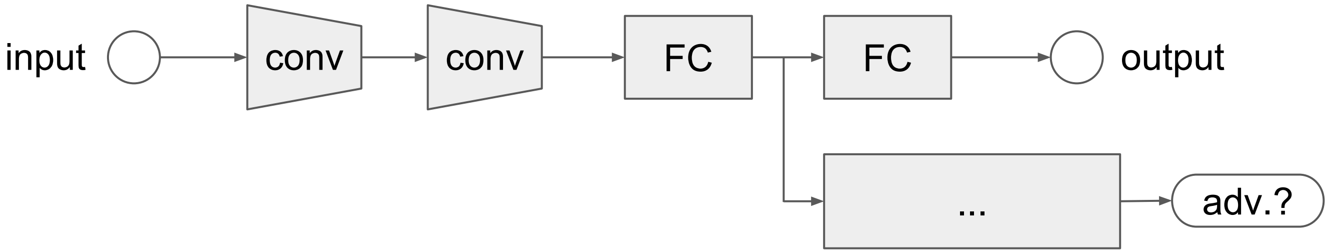
Three unrelated detectors

1. A separate network that distinguishes benign and adversarial images.
   GONG, Z., WANG, W., AND KU, W.-S. Adversarial and clean data are not twins. arXiv preprint arXiv:1704.04960 (2017).
2. The above, but using **convolution filtered images** from within the model, instead of input images.
   METZEN, J. H., GENEWEIN, T., FISCHER, V., AND BISCHOFF, B. On detecting adversarial perturbations. 5th International Conference on Learning Representations (ICLR) (2017).



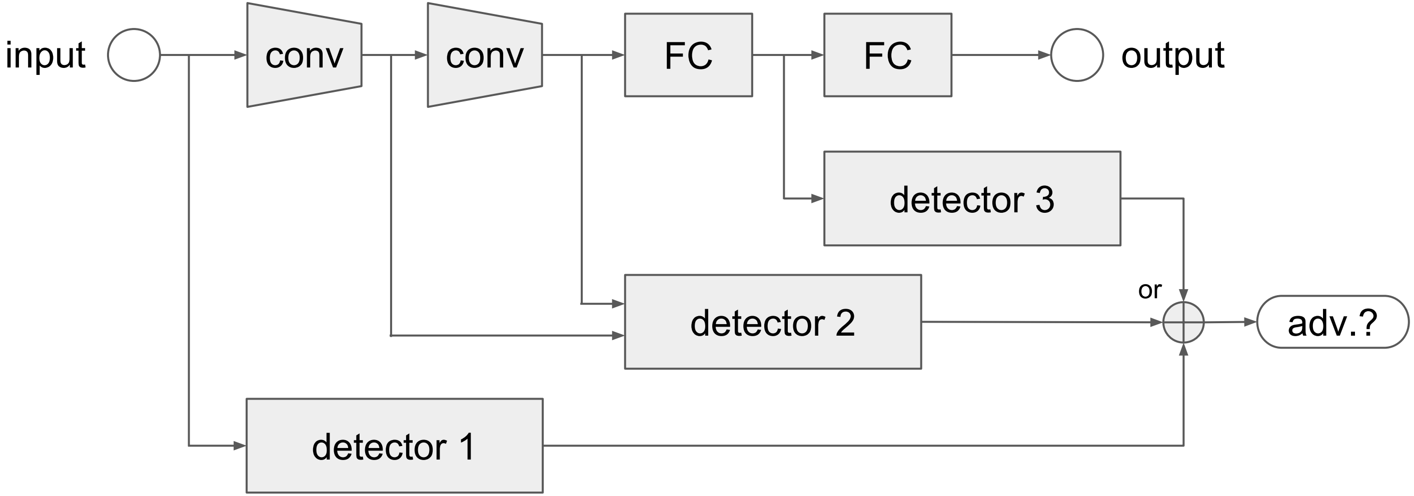
Three unrelated detectors

1. A separate network that distinguishes benign and adversarial images.
   GONG, Z., WANG, W., AND KU, W.-S. Adversarial and clean data are not twins. arXiv preprint arXiv:1704.04960 (2017).
2. The above, but using **convolution filtered images** from within the model, instead of input images.
   METZEN, J. H., GENEWEIN, T., FISCHER, V., AND BISCHOFF, B. On detecting adversarial perturbations. 5th International Conference on Learning Representations (ICLR) (2017).
3. Density estimate using Gaussian kernels, on the final hidden layer of the model.
   FEINMAN, R., CURTIN, R. R., SHINTRE, S., AND GARDNER, A. B. Detecting adversarial samples from artifacts. arXiv preprint arXiv:1703.00410 (2017).



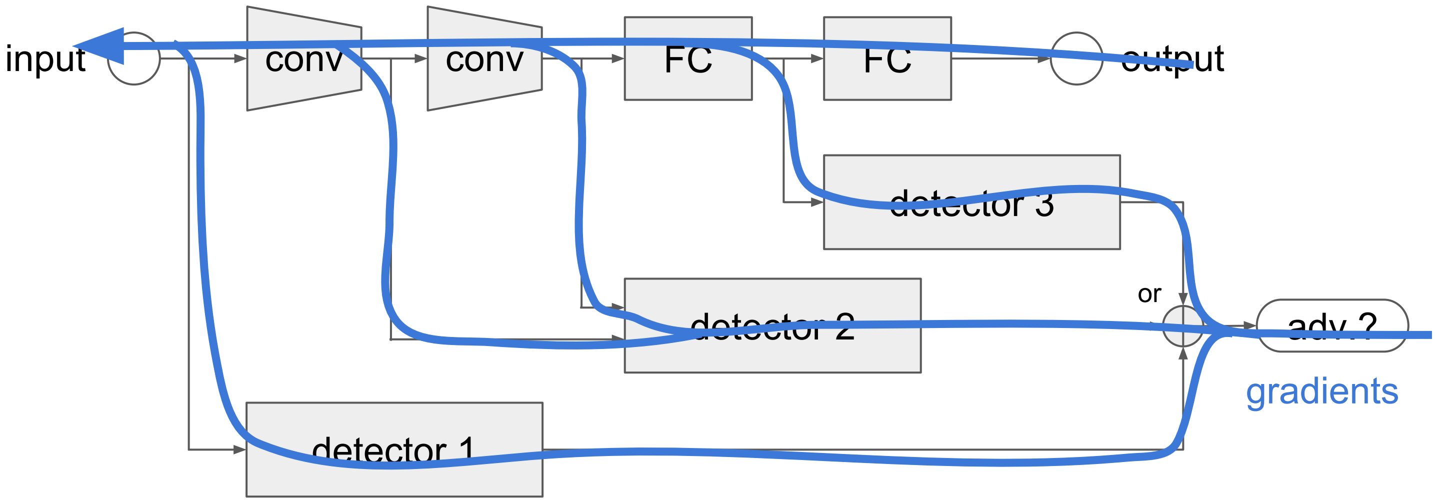
Ensemble: three unrelated detectors

If any of the three detect adversarial, system outputs adversarial.



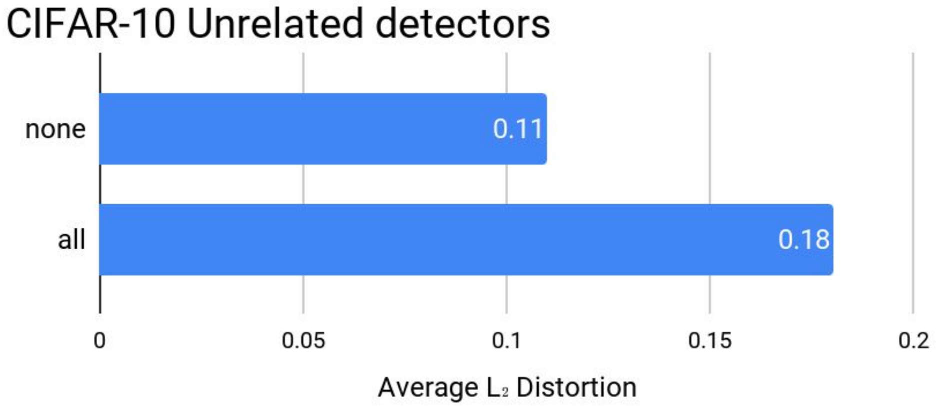
Unrelated detectors: attack

Fully differentiable system. Again, previous approaches are directly applicable.



Unrelated detectors: attack

100% success rate, imperceptible perturbations on CIFAR-10



Are ensemble defenses stronger?

Stronger! 	Not much stronger 	Weaker? 
	Feature squeezing (MNIST) +23%	
		Feature squeezing (CIFAR-10) −36%
	Specialists+1 (MNIST) +6%	
	Unrelated detectors (CIFAR-10) +60%	

Outline

Background: neural networks and adversarial examples

Defenses against adversarial examples

Ensemble defenses case studies

- Feature squeezing
- Specialists+1
- Unrelated detectors

Conclusion

Are ensemble defenses stronger?

Stronger! 

Not much stronger 

Feature squeezing (MNIST) +23%

Specialists+1 (MNIST) +6%

Unrelated detectors (CIFAR-10) +60%

Weaker? 

Feature squeezing (CIFAR-10) −36%

Are ensemble defenses stronger?

Not these ones:

- Ensembles with parts designed to work together
  - Feature squeezing
  - Specialists+1
- Unrelated detectors
  - Gong et al., Metzen et al., and Feinman et al.

Combining defenses does **not** guarantee
that the ensemble will be a stronger defense.

Conclusions

Combining defenses does **not** guarantee
that the ensemble will be a stronger defense.

Lessons:

1. Evaluate proposed defenses against **strong attacks**.
   FGSM is fast, but other methods may succeed where FGSM fails.
2. Evaluate proposed defenses against **adaptive adversaries**.
   Common assumption in security community, that attacker knows about defense, would be useful in adversarial examples research.